CASIA-CASSIL: a Chinese Telephone Conversation Corpus in Real Scenarios with Multi-leveled Annotation

Keyan Zhou[1], Aijun Li[2], Zhigang Yin[2], Chengqing Zong[1]
[1]NLPR, Institute of Automation, Chinese Academy of Sciences, Beijing, 100190
[2]Institute of Linguistics, Chinese Academy of Social Sciences, Beijing, 100732
E-mail: [1]{kyzhou, cqzong}@nlpr.ia.ac.cn, [2]{liaj, yinzg}@cass.org.cn

Abstract

CASIA-CASSIL is a large-scale corpus base of Chinese human-human naturally-occurring telephone conversations in restricted domains. The first edition consists of 792 90-second conversations belonging to tourism domain, which are selected from 7,639 spontaneous telephone recordings in real scenarios. The corpus is now being annotated with wide range of linguistic and paralinguistic information in multi-levels. The annotations include Turns, Speaker Gender, Orthographic Transcription, Chinese Syllable, Chinese Phonetic Transcription, Prosodic Boundary, Stress of Sentence, Non-Speech Sounds, Voice Quality, Topic, Dialog-act and Adjacency Pairs, Ill-formedness, and Expressive Emotion as well, 13 levels in total. The abundant annotation will be effective especially for studying Chinese spoken language phenomena. This paper describes the whole process to build the conversation corpus, including collecting and selecting the original data, and the follow-up process such as transcribing, annotating, and so on. CASIA-CASSIL is being extended to a large scale corpus base of annotated Chinese dialogs for spoken Chinese study.

1. Introduction

In order to improve the performance of systems in spoken language processing, mining and utilizing complex discourse phenomena are of paramount importance. Currently, data driven or machine learning technology will be benefited from the large-scale conversation corpus with rich phonetic, linguistic and paralinguistic annotation. In the last few decades, several English conversation corpora have been published, such as Switchboard-DAMSL (Jurafsky *et al.*, 1997) of telephone conversations, the ICSI Meeting Corpus (Janin *et al.*, 2003) and the AMI Meeting Corpus (Carletta *et al.*, 2006) of natural meetings. Annotated with abundant discourse information including dialog-acts (DAs), adjacency pairs (APs), topics, etc., these corpora greatly promote researches on English spoken discourse analysis, as well as various applications such as spoken language translation, speech recognition, spoken dialog system, and summarization as well. Meanwhile, few researches on spoken Chinese discourse have been reported, implying that such an annotated corpus of Chinese dialogs is unavailable.

CASIA-CASSIL, a large-scale corpus of Chinese spontaneous telephone conversations in tourism domain, is now being built as a fundamental corpus for study on spoken Chinese phenomena. To develop the first edition of CASIA-CASSIL, we have collected a large number of spontaneous telephone recordings up to the present. After a strict selection, only a minority of dialogs remains, which are with good voice-quality, enough turns and strictly belong to required domains. These selected dialogs are then transcribed and now being annotated with multi-leveled information, including Turns, Speaker Gender, Orthographic Transcription, Chinese Syllable, Chinese Phonetic Transcription, Prosodic Boundary, The Stress of the Sentence, Non-Speech Sounds, Voice Quality, Topic, Dialog-act and Adjacency Pairs, Ill-formedness, and Expressive Emotion.

The remainder of this paper is organized as follows. Section 2 introduces how we collect and select the original data, and gives some statistics of corpus. Section 3 describes the annotation conventions or guidelines briefly. Section 4 presents some details in annotation process, and shows an annotated example. Finally, we give concluding remarks in Section 5. In addition, Appendix A, B, and C give descriptions of annotated tags. Figure 1 gives an annotated example.

2. Data Collection and Selection

2.1 Audio Data

Restricted in tourism domain, in the first edition, we collected numbers of telephone recordings in the following four kinds of guest service centers: hotel, restaurant, airport, and travel agency. Besides, recordings are collected for taxi server when drivers have phone calls with passengers[1] . The audio sampling resolution and audio sampling rate of original recordings are 8 bits and 8 kHz.

The work on data collection lasted for about 2 years and will be continued. Some statistics of the collected data in each scenario are shown in Table 1. Although we collected numbers of recording data, only about 10% of them are selected.

2.2 Data Selection

The original audio data are first roughly transcribed manually. Data selection process is based both on transcription and original audio data. There are three criterions to judge whether a dialog is eligible or not.

1) **Quality of the recording**: the environment in many scenarios is noisy, which leads to a low quality audio record, especially in restaurant and airport. If the noise is

[1] We properly processed the problems on privacy and there is no problem on copyright.

Num of dialogues / Scenarios	Collected	Selected
Hotel	484	206
Restaurant	2,179	263
Airport	1,654	323
Travel Agency	3,032	0
Taxi	290	0
Overall	7,639	792

Table 1: Amount of Collected Data and Selected Data

Scenarios / Data amount	Hotel	Restaurant	Airport	Overall	Average
Dialogs	206	263	323	792	--
Turns	3,676	4,389	4,993	13,058	16.5
Sentences	7,352	8,778	9,986	26,116	33.0
Characters	78,950	85,491	110,135	274,576	10.5
Words	57,800	44,112	78,368	180,280	6.9

Table 2: Statistic of Selected Data

so strong that it covers up speaker's voice, it is difficult for annotators to label phonetic information. The information loss will block the following annotation process.

2) **Length of a dialog**: 10 or more turns are required. Note that our corpus is restricted to dialogs between two speakers. Once a speaker changes during the conversation, the dialog will be discarded or divided into two separated ones considering the integrality of divided dialogs. If a dialog contains less than 10 turns, it will be discarded.

3) **Content of a dialog**: since we restrict domain as tourism, we expect to collect conversations between a client and the service, instead of conversations between colleagues or outlying chats. Unfortunately, majority dialogs collected in travel agency and taxi are failed.

Thus it can be seen that compared with text corpus, collection of spoken corpus has more difficulties. Moreover, the spoken corpus selection is of huge workload. Especially for spontaneous conversations in real scenario, the circumstance and content are really beyond our control, so strict selection is essential to get the high quality corpus.

2.3 Statistic of selected data

After strict selection, we get 792 dialogs of three scenarios in total. The average length of a dialog is about 90 seconds. All the selected transcriptions are further manually corrected and cut into turns by professional annotators. Table 2 gives some statistics of the corpus. In average, a dialog contains 16.5 turns, 33 sentences. The average sentence length is 10.5 Chinese characters, 6.9 Chinese words. We get word segmentation using ICTCLAS Tagger[2].

3. Annotation Convention

The annotation is designed as a multi-leveled framework based on previous annotation systems (Li *et al.*, 2000; Li *et al.*, 2001; Li, 2002; Li and Zu, 2006). Each level is time-aligned to the audio data. Concretely speaking, each level is defined as follows.

1) **Turn (turn)**: to take count of speaker changing in the conversation.
2) **Speaker Gender (spk)**: male or female.
3) **Orthographic Transcription (HZ)**: manually corrected text with word segmentation and POS information. Specially, status of speaker is represented as service (speaker A) or client (speaker B).
4) **Chinese Syllable (PY):** the canonical syllable in Pinyin. The boundaries of syllable segments are manually divided based on audio data strictly.
5) **Chinese phonetic transcription (SY)**: including initial and final, sound change annotations, and segmentation. Especially, there are various kinds of dialects in Chinese. Most of the speakers are bilingual speakers in dialects or regional accent Mandarin. Some of regional accent and misspeaking are also represented in this level.
6) **Prosodic boundary (BI)**: prosodic structures including prosodic phrase and intonation phrase and turn boundaries.
7) **The stress of the sentence (ST)**: the stress of each intonation phrase.
8) **Voice quality (VQ)**: to describe the phonation information of the speakers, such as falsetto, whisper, creaky, etc.
9) **Non-speech sounds (MIS)**: to note nonlinguistic phenomena including non-verbal background noise such as ringing, door opening sounds, and breath, coughing, cry, laugh, and so on.
10) **Topic (TP)**: an open set includes opening, closing, inquire, advice, request, reservation, and others. The definition is described in Appendix A. It will be enriched while annotating process.
11) **DA and AP (dialogact)**: similar to DA definition in ICSI-MRDA corpus (Dhillon *et al.*, 2004), the unit of DA is utterance. A dialog will be segmented into utterances before being labeled with DA tags.
    There are two levels of DA tags: general tags (9 labels) which represent the basic form of an utterance (e.g., statement, question, etc.), and appended specific tags (36 labels) which represent the function or characteristics of an utterance. Specially, considering the integrality of utterance when turn changes, a tag set called interruption is proposed, which contains 3 tags (abandoned, interrupted, and indecipherable). Each utterance needs one general tag; meanwhile, it might contain one or more specific tags. General tag and specific tag are connected by symbol'^'. If the utterance is not integrated, an interruption tag will be appended. Interruption tag follows general tag and specific tag with symbol '.'. The definition of each DA tag is described in Appendix B.

[2] http://ictclas.org/

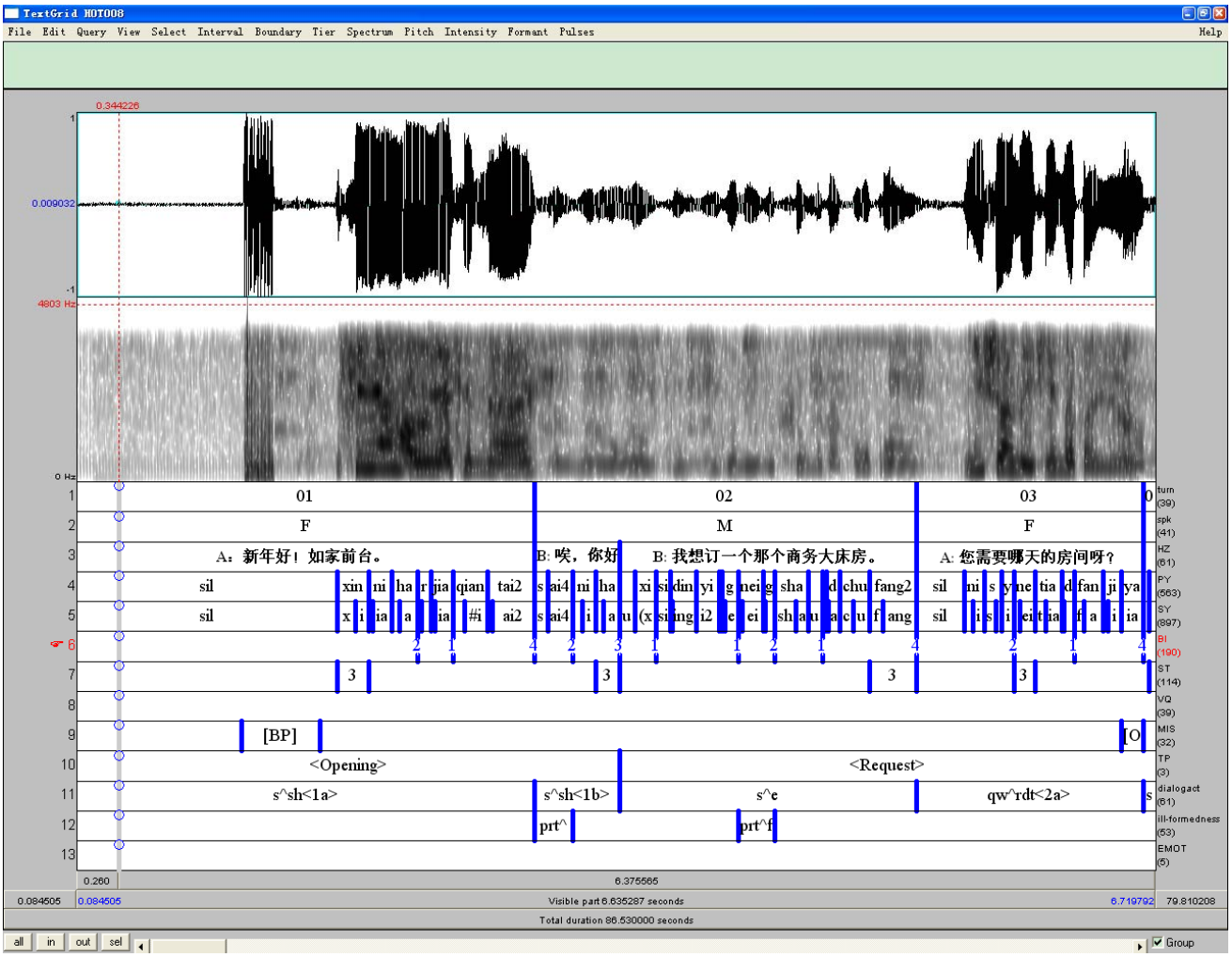

Figure 1: The Annotation Interface using Praat
(three panels from top to bottom are for waveform, spectrogram and 13 annotation layers. Annotation file with detailed statement is shown in Appendix D.)

APs are paired utterances, defined as one kind of sociolinguistic facts about conversation structure, which is a reflection of dialog structure (Levinson, 1983). An AP consists of two parts produced by different speaker (Dhillon *et al*., 2004). In our work, AP contains the following relationships of the utterances: question-answer, greeting-greeting, offer-acceptance, and apology-downplay.

12) **Ill-formedness (ill-formedness)**: Since the dialogs are naturally-occurred, ill-formedness utterances are unavoidable. We give three basic categories to describe part of ill-formedness phenomenon, and divide the three categories into 13 patterns in total. Details are shown in Appendix C.
13) **Expressive emotion (EMOT)**: there are 70 kinds of expressive representations, such as happy, appreciate, scared, worry, surprise, and so on. Each expressive emotion has two grades to further measure the degree of expressiveness.

4. Software and Annotation Consistency

Our annotation software is Praat [3], a well-known, well-maintained and widely used tool for speech annotating, analyzing, synthesizing, and manipulating. Figure 1 shows the annotation interface using Praat.

In praat, prosodic boundary (BI) is a TextTier object, which contains a series of time points. The other levels are IntervalTier objects, which contain a series of contiguous intervals in time. The textual annotations are saved in TextGrid format. We give a simplified example in Appendix D.

There are three professional annotators in charge of transcription and annotation. First, the recorded spontaneous dialogues are selected manually, only the qualified episodes were transcribed into orthographic texts (Chinese characters) manually and then automatically chopped into short utterances. Chinese characters were transcribed automatically into orthographic Pinyin as well. After this, multi-layer's annotations including segmental, prosodic, linguistic, and paralinguistic and pragmatic information were done in praat according to the conventions as described above.

To keep the consistency degree of annotation, the three annotators discussed a lot during the first stage with the authors and checked the annotations of the first 100 dialogs among each other. Through the checking, the

[3] http://www.praat.org

annotation guidelines were consummated, meanwhile, the consistency was improved. During the whole procedure, the annotated results were selected randomly and checked by the first author. After annotation work finished, all the labels will be checked throughout.

5. Conclusion

In this paper, we introduce a Chinese conversational corpus CASIA-CASSIL which is developed based on huge amount of telephone recordings occurred in natural scenarios. After strict selection, 792 dialogs restricted in tourism domain constitute the first edition of CASIA-CASSIL. Although the corpus is still under construction, it is expected to be the first large-scale Chinese spontaneous conversation corpus. The corpus will be annotated with multi-leveled labels in a wide range of linguistic and phonetic information. The corpus contains not only phonetic annotation, but also semantic, emotion and discourse information. We believe it will be widely used in spontaneous speech and discourse analysis and applied in spoken language translation system, speech recognition, spoken dialog system, dialog summarization and any other application systems.

6. Acknowledgements

We would like to thank Bingye Wang for his work on data collecting, Jiao Lei, Jing Chen, Yu Yu, Mengqun Zhai, Yanmin Dong, and Liqin Wang for their work on original data transcription, Ting Fang, Guohong Tian, Hongli Liang for their work on data selection, transcription correction, and annotation. We specially thank Ting Fang for her suggestion on annotation guideline.

The research work described in this paper has been partially funded by the Natural Science Foundation of China under Grant No. 60975053 and 90820303, and also supported by the National Key Technology R&D Program under Grant No. 2006BAH03B02, and CASS Key Lab project.

7. References

Carletta, J., S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraajj, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner. (2006). The AMI Meeting Corpus: A Pre-Announcement. In Steve Renals and Samy Bengio, editors.

Dhillon, R., S. Bhagat, H. Carvey, and E. Shriberg. (2004). Meeting Recorder Project: Dialog-act Labeling Guide. ICSI Technical Report TR-04-002, International Computer Science Institute.

Janin, A., D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, C. Wooters. (2003). The ICSI Meeting Corpus. In *Proceedings of the 28st International Conference on Acoustics, Speech, and Signal Processing*, Hong Kong.

Jurafsky, D., L. Shriberg, and D. Biasca. (1997). Switchboard SWBD-DAMSL Labeling Project Coder's Manual, Draft 13. Technical Report 97-02, University of Colorado Institute of Cognitive Science.

Levinson, Stephen C. (1983). *Pragmatics*. Cambridge: Cambridge University Press.

Li, A., F. Zheng, W. Byrne, *et al.*, (2000). Cass: "A Phonetically Transcribed Corpus of Mandarin Spontaneous", *ICSLP'2000*.

Li, A., B. Xu, C. Aong, *et al.* (2001). A Spontaneous Conversation Corpus CADCC, *Oriental COCOCSDA'2001*, Korea.

Li, A. (2002). Chinese Prosody and Prosodic Labeling of Spontaneous Speech. In B.Bel and I. Marlin (eds), *Proceedings of the Speech Prosody 2002 Conference*. Aix-en-Provence, France, 2002: 39-46.

Li, A., Y. Zu. (2006). Corpus Design and Annotation for Speech Synthesis and Recognition, as a chapter in Advances in *Chinese Spoken Language Processing*, edited by Chin-Hui Lee, Haizhou Li, Lin-shan Lee, Ren-Hua Wang, Qiang Huo, World Scientific Publishing Co. Pte. Ltd., Singapore. pp. 243-268.

Appendix A. Topic Description.

Topic	Description
Opening	Say hello, welcome.
Closing	Thanks and goodbye.
Inquire	Inquire specific information, such as name, telephone number, address, time, and so on.
Advice	Opinion and advice.
Request	Request for room service, membership discount, etc.
Reservation	For hotel reservation, restaurant reservation, ticket booking, etc.
Others	Will be enriched while annotating process.

Appendix B. Dialog-Act Definition.

I. General Tags

Tag	Description
s	Statement
qy	Y/N Question
qw	Wh-Question
qr	Or Question
qrr	Or Clause After Y/N Question
qo	Open-end Question
qh	Rhetorical Question
is	Imperative Sentence
es	Exclamatory Sentence

II. Interruptions

Tag	Description
%-	Interrupted
%--	Abandoned
%	Indecipherable

III. Specific Tags

Group		Tag	Description
1. Responses	Positive	aa	Accept
		aap	Partial Accept
		na	Affirmative Answer
	Nagative	ar	Reject

		arp	Partial Reject
		nd	Dis-preferred Answer
		ng	Negative Answer
	uncertain	am	Maybe
		no	No Knowledge
2. Action Motivators		co	Command
		cs	Suggestion
		cc	Commitment
		so	Soliloquy
3. Checks		f	“Follow Me”
		br	Repetition Request
		bu	Understanding Check
		b	Backchannel
4. Restated Information	Repetition	r	Repeat
		m	Mimic
		bs	Summary
	Correction	bc	Correct Misspeaking
		bsc	Self-Correct Misspeaking
		bsa	Self-affirm
5. Supportive Functions		df	Defending /Explanation
		e	Elaboration
		2	Collaborative Completion
6. Politeness Mechanisms		bd	Downplayer
		by	Sympathy
		fa	Apology
		ft	Thanks
		fw	Welcome
		sh	Say Hello
		bye	Bye
7. Request		raf	Request Affirmation
		rdt	Request Details
		rsg	Request Suggestion

Appendix C. Ill-formedness Patterns.

Pattern	Description	Tag
1. Parenthesis	Floor Grabber	prt^fg
	Floor Holder	prt^fh
	Hold	prt^h
	Third Party Talk	prt^t3
2. Overlapping	Entire Overlapped	rpt^cf
	Partly Overlapped	rpt^xd
	Different Expression but with the Same Meaning	rpt^yz
	Misspeaking Correction	rpt^fd
	Complex	rpt^fz
3. Disorder	Subject Disordered	ovt^zh
	Object Disordered	ovt^bq
	Modifier Disordered	ovt^xh
	Complex	ovt^qt

Appendix D. A TextGrid file of an Annotated Dialog

```
File type = "ooTextFile"
Object class = "TextGrid"
xmin = 0                    /*starting time*/
xmax = 86.53                /*end time*/
tiers? <exists>
size = 13                   /*13 levels*/
item []:
   item [1]:                /*level 1: Turn*/
      class = "IntervalTier" /*contiguous intervals in time*/
      name = "spk"
      xmin = 0
      xmax = 86.53
      intervals: size = 39    /*total number of labels in level 1*/
      intervals [1]:          /*an annotation*/
         xmin = 0             /*starting time*/
         xmax = 2.8990665757540235  /*end time*/
         text = "01"          /*label*/
      intervals [2]:
         xmin = 2.8990665757540235
         xmax = 5.246733925149825
         text = "02"
      intervals [3]:
         xmin = 5.246733925149825
         xmax = 6.643020539909597
         text = "03"
      intervals [4]:
         ……

   item [2]:                /*level 2: Speaker Gender*/
      class = "IntervalTier"
      name = "spk"
      xmin = 0
      xmax = 86.53
      intervals: size = 41
      intervals [1]:
         xmin = 0
         xmax = 2.8990665757540235
         text = "F"
      intervals [2]:
         xmin = 2.8990665757540235
         xmax = 5.246733925149825
         text = "M"
      intervals [3]:
         xmin = 5.246733925149825
         xmax = 6.643020539909597
         text = "F"
      intervals [4]:
         ……

   item [3]:                /*level 3: Orthographic Transcription*/
      class = "IntervalTier"
      name = "HZ"
      xmin = 0
      xmax = 86.53
      intervals: size = 61
      intervals [1]:
         xmin = 0
         xmax = 2.8990665757540235
         text = "A：新年好！如家前台。"
      intervals [2]:
         xmin = 2.8990665757540235
         xmax = 3.4242236544782725
         text = "B：唉，你好！"
      intervals [3]:
         xmin = 3.4242236544782725
         xmax = 5.246733925149825
         text = "B：我想订一个那个商务大床房。"
      intervals [4]:
         ……
```

```
item [4]:                /*level 4: Chinese Syllable*/
    class = "IntervalTier"
    name = "PY"
    xmin = 0
    xmax = 86.53
    intervals: size = 563
    intervals [1]:
        xmin = 0
        xmax = 1.6863598483498574
        text = "sil"
    intervals [2]:
        xmin = 1.6863598483498574
        xmax = 1.882377605361048
        text = "xin1"
    intervals [3]:
        xmin = 1.882377605361048
        xmax = 2.024093145902922
        text = "nian2"
    intervals [4]:
        ……

item [5]:              /*level 5: Chinese Phonetic Transcription*/
    class = "IntervalTier"
    name = "SY"
    xmin = 0
    xmax = 86.53
    intervals: size = 897
    intervals [1]:
        xmin = 0
        xmax = 1.6863598483498574
        text = "sil"
    intervals [2]:
        xmin = 1.6863598483498574
        xmax = 1.798937614200879
        text = "x"
    intervals [3]:
        xmin = 1.798937614200879
        xmax = 1.882377605361048
        text = "in1"
    intervals [4]:
        ……

item [6]:                  /*level 6: Prosodic Boundary*/
    class = "TextTier"   /* time points */
    name = "BI"
    xmin = 0
    xmax = 86.53
    points: size = 190
    points [1]:
        time = 2.1803775737902225
        mark = "2"
    points [2]:
        time = 2.398910883971617
        mark = "1"
    points [3]:
        time = 2.8990665757540235
        mark = "4"
    points [4]:
        ……

item [7]:                  /*level 7: the Stress of the Sentence*/
    class = "IntervalTier"
    name = "ST"
    xmin = 0
    xmax = 86.53
    intervals: size = 114
    intervals [1]:
        xmin = 0
        xmax = 1.6863598483498574
        text = ""
    intervals [2]:
        xmin = 1.6863598483498574
        xmax = 1.882377605361048
        text = "3"
    intervals [3]:
        xmin = 1.882377605361048
        xmax = 3.2758858924157503
        text = ""
    intervals [4]:
        ……

item [8]:                  /*level 8: Voice Quality*/
    class = "IntervalTier"
    name = "VQ"
    xmin = 0
    xmax = 86.53
    intervals: size = 39
    intervals [1]:
        xmin = 0
        xmax = 7.163546520853924
        text = ""
    intervals [2]:
        xmin = 7.163546520853924
        xmax = 7.301584146835802
        text = "CR"
    intervals [3]:
        xmin = 7.301584146835802
        xmax = 11.469061958630712
        text = ""
    intervals [4]:
        ……

item [9]:                  /*level 9: Non-Speech Sounds*/
    class = "IntervalTier"
    name = "MIS"
    xmin = 0
    xmax = 86.53
    intervals: size = 32
    intervals [1]:
        xmin = 0
        xmax = 1.1005389235900631
        text = ""
    intervals [2]:
        xmin = 1.1005389235900631
        xmax = 1.5828631757326725
        text = "[BP]"
    intervals [3]:
        xmin = 1.5828631757326725
        xmax = 6.509471607300766
        text = ""
    intervals [4]:
        ……

item [10]:                 /*level 10: Topic*/
    class = "IntervalTier"
    name = "TP"
    xmin = 0
    xmax = 86.53
    intervals: size = 3
    intervals [1]:
        xmin = 0
        xmax = 3.4242236544782725
        text = "<Opening>"
    intervals [2]:
        xmin = 3.4242236544782725
        xmax = 79.94247962537013
        text = "<Request>"
    intervals [3]:
        xmin = 79.94247962537013
        xmax = 86.53
        text = "Closing"
```

```
item [11]:                  /*level 11: DA and AP*/
    class = "IntervalTier"
    name = "dialogact"
    xmin = 0
    xmax = 86.53
    intervals: size = 61
    intervals [1]:
        xmin = 0
        xmax = 2.8990665757540235
        text = "s^sh<1a>"
    intervals [2]:
        xmin = 2.8990665757540235
        xmax = 3.4242236544782725
        text = "s^sh<1b>"
    intervals [3]:
        xmin = 3.4242236544782725
        xmax = 5.246733925149825
        text = "s^e"
    intervals [4]:
        ……

item [12]:                  /*level 12: Ill-Fromedness*/
    class = "IntervalTier"
    name = "ill-formedness"
    xmin = 0
    xmax = 86.53
    intervals: size = 53
    intervals [1]:
        xmin = 0
        xmax = 2.8990665757540235
        text = ""
    intervals [2]:
        xmin = 2.8990665757540235
        xmax = 3.1368192404821356
        text = "prt^h"
    intervals [3]:
        xmin = 3.1368192404821356
        xmax = 4.1505558501482796
        text = ""
    intervals [4]:
        ……

item [13]:                  /*level 13: Expressive Emotion*/
    class = "IntervalTier"
    name = "EMOT"
    xmin = 0
    xmax = 86.53
    intervals: size = 5
    intervals [1]:
        xmin = 0
        xmax = 11.590737559844415
        text = ""
    intervals [2]:
        xmin = 11.590737559844415
        xmax = 13.529001804723572
        text = "yih1"
    intervals [3]:
        xmin = 13.529001804723572
        xmax = 17.08285908326869
        text = ""
    intervals [4]:
        ……
```